35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

EpiDiff: entropy-based quantitative identification of differential epigenetic modification regions from epigenomes*

Yan Zhang, Jangzhong Su, Di Yu, Qiong Wu and Haidan Yan

Abstract—**Genome-wide epigenetic modification dynamics, including DNA methylation and chromatin modification, are involved in biological processes such as development, aging, and disease. Quantitative identification of differential epigenetic modification regions (DEMRs) from various temporal and spatial epigenomes is a crucial step towards investigating the relationship between epigenotype and phenotype. Here, we describe EpiDiff (http://bioinfo.hrbmu.edu.cn/epidiff/), an integrated software platform that supports quantification of epigenetic difference and identification of DEMRs by Shannon entropy. Two main modules, quantitative differential chromatin modification region (QDCMR) and quantitative differentially methylated region (QDMR) are provided for bioinformatic analysis of chromatin modifications and DNA methylation data, respectively. The third module, quantitative differential expressed gene (QDEG), can be used to identify differentially expressed genes. The platform-free and species-free nature of EpiDiff makes it potentially applicable to a wide variety of epigenomes at an unprecedented scale and resolution. The graphical user interface provides biologists with a practicable and reliable way to analyze and visualize epigenetic difference.**

I. Introduction

Epigenetic modifications play critical roles in the regulation of gene expression and chromatin remodeling. Promoter hypermethylation can suppress gene transcription directly by inhibiting the binding of transcription factors to their target sites (1). Chromatin modifications, including histone variants and their post-translational modifications, also play an important role in regulating gene expression (2). And aberrant epigenetic changes in these regions are involved in disease processes (3). Differential epigenetic modification regions (DEMRs), as genomic regions with different epigenetic statuses among multiple samples (tissues, cells, individuals or others)(4,5.6), are regarded as possible functional regions involved in gene regulation. Dynamic epigenetic modifications in DEMRs are fundamental to the regulation of many cellular processes, including cell development, differentiation, X-chromosome inactivation and genomic imprinting (7,8).

*Research supported by the Scientific Research Fund of Heilongjiang Provincial Education Department (12511272).

Yan Zhang is with the College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China (Tel/Fax: +86 451 8666 7543; Email: yanyou1225@163.com).

Jianzhong Su and Haidan Yan are with the College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China (Email: jianzhongsu@yahoo.cn, haidanyan@yahoo.cn).

Di Yu is with the Heilongjiang University, Harbin, 150081, China (Email: tyozhang@ems.hrbmu.edu.cn).

Qiong Wu is with School of Life Science and Biotechnology, Harbin Institute of Technology, Harbin, 150001, China(Email: kigo@hit.edu.cn).

High-throughput experimental techniques using microarrays and next-generation sequencing are providing epigenomic data on an unprecedented scale. Several techniques, such as RRBS (9) and MethylC-Seq(10), have been developed for profiling DNA methylation patterns across various cells or tissues. In most of these techniques, the original or pretreated DNA methylation status is represented by continuous values, with a measurement scale from 0 to 1 (11). Recently, chromatin immunoprecipitation (ChIP) followed by microarray hybridization (ChIP-Chip) or high-throughput sequencing (ChIP-Seq) has been widely used for genome-wide profiling of chromatin modifications and DNA-binding proteins (12). The unprecedented scale and precision of epigenomic data have enabled the quantitative analysis of differential epigenetic status in gene regulation in various biological processes. Thus, effective computational tools for mining epigenetic differences are crucial for uncovering biological mechanisms of development, aging, and disease.

Over recent years, considerable efforts have been made in the identification of DEMRs from high throughput epigenome data. Both statistics-based and counting-based methods have been used for identification of DMRs cross multiple cells/tissues. In our previous work, we developed an entropy-based quantitative approach, QDMR, for quantification of methylation difference and identification of DMRs across multiple samples from various methylomes (13). There are also some methods for analysis of differential chromatin modification. Based on hidden Markov model (HMM), Xu et al. proposed an approach, ChIPDiff, for the genome-wide identification of differential histone modification sites from ChIP-Seq data (14). RSEG, developed by Song et al., identifies dispersed epigenomic domains from ChIP-Seq data, and can be used to identify DCMRs with a three-state HMM (15). However, both of these methods depend on certain distributions which ChIP-Seq data may not always follow because of different ChIP-Seq data preprocessing methods. In addition, both of these methods are only applicable to the identification of DCMRs between two samples.

In this context, we describe EpiDiff, based Shannon entropy (16), which is a quantitative measure of difference and uncertainty in a data set and has been widely applied in quantitative biology. The analysis method used in our previous tool QDMR for DNA methylation is extended to analysis of chromatin modification, trans-acting factor binding sites and gene expression data by appropriate adjustments. EpiDiff is a user-friendly integrated software tool that supports quantification of epigenetic difference and identification of DEMRs, including DMRs and DCMRs. In addition, this tool also can be used to identify DTAFSs and differentially

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

expressed genes (DEGs). EpiDiff provides effective tools for the high-throughput identification of functional regions involved in epigenetic regulation.

I. Program overview

EpiDiff facilitates genome-wide quantitative comparison and visualization of epigenetic modifications among multiple samples. The tool takes epigenomic data as the input, and produces tabular and graphical output of the quantified epigenetic difference, differential regions, sample specificity, and genome information at the UCSC Genome Browser.

EpiDiff is a Java-based program that can be run on computers with a recent version of the Java Virtual Machine installed. This tool can be run as a self-installing distribution directly via Java Web Start, and as a local standalone installation package for offline computers. The module-based friendly user interface of EpiDiff supports the comprehensive analysis of a variety of high-throughput epigenetic data in multiple genome regions across multiple samples. Three modules (QDCMR, QDMR, and QDEG) are provided for bioinformatic analysis of chromatin modifications and trans-acting factor binding sites data, DNA methylation data, and gene expression data, respectively (Figure 1A). All modules can be accessed from within each module. For each module, a typical analysis consists of five phases: (i) data import; (ii) difference quantification; (iii) differential region identification; (iv) specificity measurement; and (v) data visualization and export (Figure 1B).

II. Data Import

Each of the three modules starts from importing data, either from the laboratory or processed by bioinformatics methods via the corresponding import interface (Figure 2A). For every import interface, two example data are provided as references. All example data can be downloaded from EpiDiff website. It is suggested that users refer to and execute the example data before import their own data. All the import interfaces of the three modules support import of processed data in txt/xls files, the format of which is shown in the corresponding example data. Information about the regions of interest should be noted before the sample data for the region is processed. Users should also ensure that there are no missing values in the import data. Users also can define the column names, species, region information columns, sample information columns, data range (only for QDMR module) by the import interfaces. Moreover, the first 20 rows of data will be shown in the data file preview window embedded in these interfaces, which enables users to preview and re-define the imported data.

A special interface is provided in the QDCMR module for importing the raw chromatin modification data by ChIP-Seq (Figure 2B). In this interface, two types of data should be imported. One is the region file which contains the regions of interest. The other data are the raw data files, including chromatin modification reads that have been aligned to the genome. These data files can be saved in .bed or txt.gz format, both of which are widely used in ChIP-Seq data. In addition, these files should be imported into the software as a file folder or a compressed file (e.g. a zip file), as shown in the example data.

III. Data processing

EpiDiff implements a data processing pipeline that is run for each region across multiple samples. The pipeline quantifies the difference among the data in all samples by entropy. Based on these entropy values, it infers which regions are differential among these samples by a threshold determined from the probability model in EpiDiff. The specificity in each sample is then measured for each differential region. More details about the key steps of the data processing pipeline are outlined in below. All of these analyses can be conveniently implemented by mouse clicks in the graphical user interface, which provides biologists with a practicable and reliable way to analyze and visualize epigenetic differences.

IV. Quantification of epigenetic difference by entropy

In this study, we selected Shannon entropy, a quantitative measure of difference and uncertainty in a data set, which has been widely applied in quantitative biology (16). Furthermore, we performed several optimizations of the algorithm to account for recurrent issues with epigenetic modification data. To equally quantify the modification difference of the regions with hyper- or hypo-modification in minor samples, a one-step Tukey's biweight is used to process the raw modification levels for each region, as Kadota et al. did in the development of the ROKU method (17). Considering the range of variation of the modification data, the entropy for each region is adjusted by a modification weight that was defined based on the ratio between the data range among samples in the region and the total data range. The higher the entropy, the larger the epigenetic modification difference across multiple samples. The entropy determined by this adapted method can accurately represent the degree of epigenetic modification difference among multiple samples. In addition to identifying differential regions, the entropy inferred by EpiDiff also can be used to quantitatively analyze the correlation between different types of epigenetic modification difference or the relationship between epigenetic modification difference and gene expression difference.

V. Identification of differential regions by threshold

Based on the quantitative modification difference, differential regions can be identified by an appropriately defined threshold. In this study, the threshold is determined by a modification probability model, in which the random biological variability among samples was modeled based on the assumption that each region exhibits an average modification level across all samples. The log base 2 of the fold change between replicate-dependent difference from the average level across replicates and the theoretical maximum range of epigenetic modification was assumed to display a normal distribution with a mean equal to zero and a standard deviation.

After setting the proper standard deviation, the mean modification intensity is sampled from the distribution of observed mean epigenetic modification intensities obtained randomly from the data submitted by user, and 5000 regions with uniform epigenetic modification across samples were modeled. The entropy values of these 5000 regions follows a normal distribution in which a threshold is determined at p = 0.05 (one-sided). This process is repeated 10 times, and the mean value of 10 thresholds is defined as the threshold for identification of differential regions across multiple samples. The regions with entropy lower than the threshold are identified as differential regions, which are widely used in the comparative genomics and epigenomics.

VI. Measurement of Sample Specificity for Differential Regions

The sample-specific modification levels are considered as the main individual factors that determine the modification differences across samples. Here, the contribution of a selected sample to the whole modification difference is reflected by the difference between the entropy across all samples and the entropy across the samples that do not include the selected sample. Thus, a positive entropy difference represents a sample-specific modification, while a negative one represents no specificity and is replaced by 0. To further identify specific hyper-modification or hypo-modification in a region, the categorical sample-specificity (CS) is defined as the product of entropy difference and sign of the difference between the modification level in the selected sample and the median modification level in all samples.

VII. Data Visualization and Export

Once each step of data processing is complete, the results are shown in the data table on the right panel of the software (Figure 3A). The data in the first row are visualized in the visualization window acquiescently. Users can click a row to view the data across samples and set the image properties by right clicking. In addition, users can double-click a row to view the region information in the UCSC Genome Browser (Figure 3B) if the information about species, chromosome, region start, and region end have been defined in the import interface. All results can be exported in several formats, including data tables and graphical plots. The visualization module allows the user to inspect the raw data pattern, distribution of differential regions on chromosomes, and genome information in UCSC Genome Browser. Moreover, the graphical output can be published in research papers to explain the analysis results clearly.

Figures and Tables

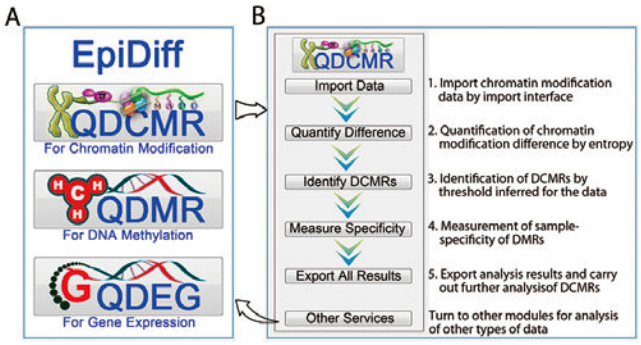

Figure 1. Overview of EpiDiff. (A) Three modules (QDCMR, QDMR, and QDEG) of EpiDiff for chromatin modification, DNA methylation, and gene expression analysis. (B) The workflow of QDCMR module.

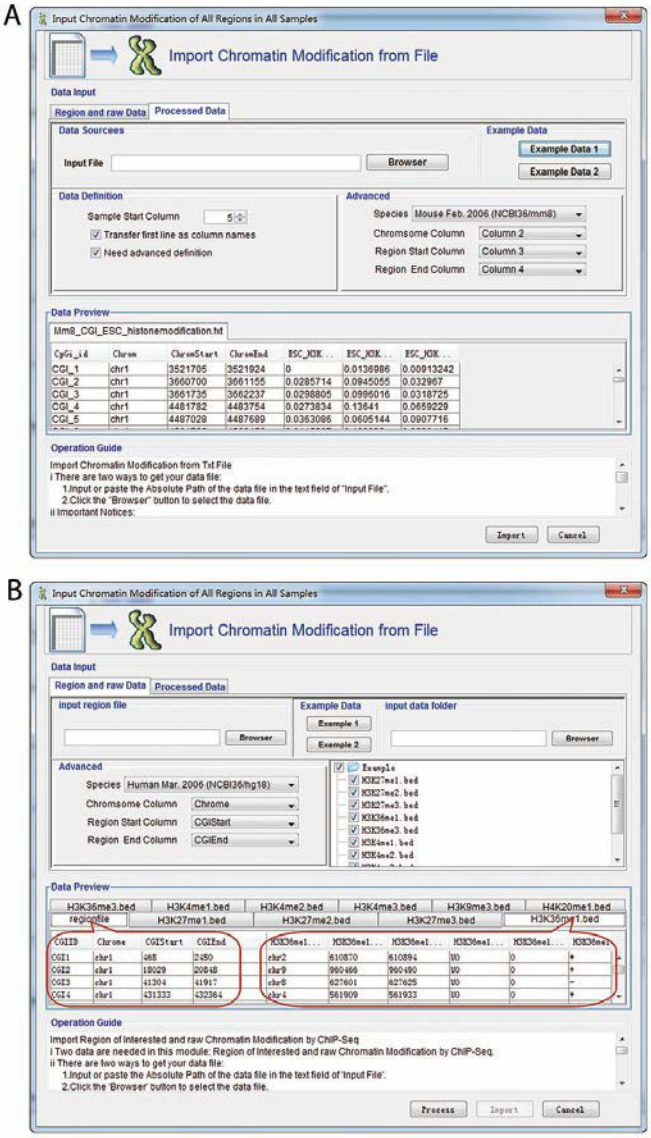

Figure 2. Data import interfaces of EpiDiff. (A) The import interface of processed data. (B) The import interface of raw chromatin modification data by ChIP-Seq in QDCMR module.

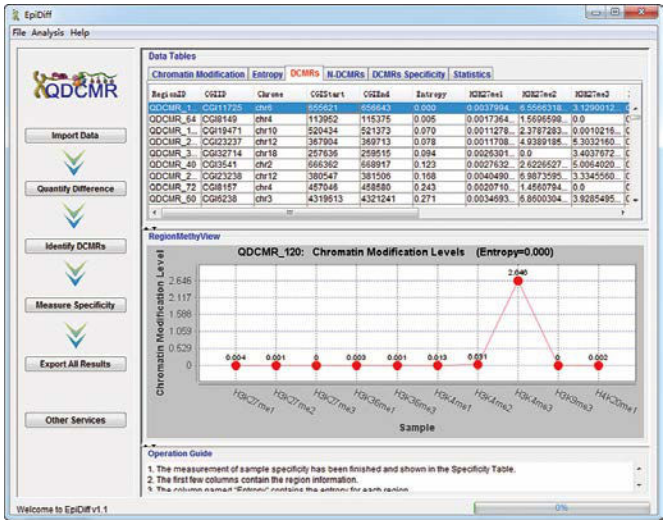

Figure 3. Data visualization and export. (A) The visual interface of EpiDiff. (B) EpiDiff provides a convenient entrance to the genome annotation in the UCSC Genome Browser of the studied region.

VIII. Conclusion

EpiDiff is a user-friendly integrated software platform, which provides comprehensive support for quantification of epigenetic difference, identification of differential regions, and measurement of sample specificity across multiple samples. The bioinformatic challenges in genome-wide quantitative analysis of epigenetic difference are addressed by a specifically optimized Shannon entropy algorithm. The platform-free and species-free nature of EpiDiff makes it potentially applicable to unprecedented scale epigenomes profiled by high-throughput experimental techniques using microarrays and next-generation sequencing. In summary, EpiDiff provides effective tools for the quantitative identification of the differential regions and the potential biomarkers involved in epigenetic regulation.

References

1. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev*, **16**, 6-21.
2. Kouzarides, T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693-705.
3. Portela, A. and Esteller, M. (2010) Epigenetic modifications and human disease. *Nat Biotechnol*, **28**, 1057-1068.
4. Rakyan, V.K., Down, T.A., Thorne, N.P., Flicek, P., Kulesha, E., Graf, S., Tomazou, E.M., Backdahl, L., Johnson, N., Herberth, M. *et al.* (2008) An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res*, **18**, 1518-1529.
5. Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat Genet*.
6. Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766-770.
7. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553-560.
8. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, **40**, 897-903.
9. Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S. and Jaenisch, R. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res*, **33**, 5868-5877.
10. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315-322.
11. Laird, P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet*, **11**, 191-203.
12. Zhou, V.W., Goren, A. and Bernstein, B.E. (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews. Genetics*, **12**, 7-18.
13. Zhang, Y., Liu, H., Lv, J., Xiao, X., Zhu, J., Liu, X., Su, J., Li, X., Wu, Q., Wang, F. *et al.* (2011) QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res*, **39**, e58.
14. Carmona, F.J. and Esteller, M. (2010) Epigenomics of human colon cancer. *Mutation research*, **693**, 53-60.
15. Claes, B., Buysschaert, I. and Lambrechts, D. (2010) Pharmaco-epigenomics: discovering therapeutic approaches and biomarkers for cancer therapy. *Heredity*, **105**, 152-160.
16. Shannon, C.E. (1963) The mathematical theory of communication. *MD Comput*, **14**, 306-317.
17. Kadota, K., Ye, J., Nakai, Y., Terada, T. and Shimizu, K. (2006) ROKU: a novel method for identification of tissue-specific genes. *BMC Bioinformatics*, **7**, 294.